Spell checking an agglutinative language: Quechua

Rios, A

Abstract: Spelling correction methods developed for languages like English usually rely on complete lists of full word forms, a requirement that cannot be met for morphologically complex languages. In this article we describe the implementation of a spell checker using finite state methods for the agglutinative language Quechua (ISO 639-3:que).

Spell Checking an Agglutinative Language: Quechua

Annette Rios

Institute for Computational Linguistics
University of Zurich
arios@ifi.uzh.ch

Abstract

Spelling correction methods developed for languages like English usually rely on complete lists of full word forms, a requirement that cannot be met for morphologically complex languages. In this article we describe the implementation of a spell checker using finite state methods for the agglutinative language Quechua (ISO 639-3:que).

Keywords: LRL, Quechua, foma, xfst, morphology, spell checking

1. Introduction

Spell checking is an important tool for the writing of texts, and almost every text processing system has a module to deal with misspelled words. The process of spelling correction consists of two tasks: In a first step, the spell checker decides whether a given word form is correct. If this is not the case, in a second step, correct word forms close to the input have to be found in order to suggest a correction.

Spelling correction methods developed for languages like English usually rely on complete lists of word forms, a requirement that cannot be met for morphologically complex languages like Quechua.

Each nominal or verbal root in Quechua may be used in thousands of possible word forms, therefore the compilation of fully fledged word lists is not feasible. A more adequate approach to capture the morphological structures of agglutinative languages are finite state techniques. Spell checkers relying on finite state methods have been described for Turkish (Oflazer, 1996), Finnish (Pirinen and Lindén, 2010) and Basque (Alegria et al., 2002).

We recently started a project about Spanish to Quechua machine translation. Due to its agglutinative structure, the first requirement for any automatic processing of Quechua are tools to handle its rich morphology. Therefore, we have implemented a morphological analyzer and generator for Quechua in *xfst*. A slight adaption of these tools for spell checking can be done with little effort. The error metric of choice is the Levenshtein distance, obtained by calculating how many basic edit operations (deletion, insertion and substitution) are necessary to convert one string into the other. In order to use the *xfst* tools for spell checking, the edit operations have to be explicitly allowed in the finite state automaton, as a consequence, for every transition in the already large automaton, three new possibilities arise: a transition can be removed, replaced or an additional transition may be inserted. The resulting *xfst* finite state automaton is huge and therefore too slow to be used in a real application. For this reason, we re-implemented the spell checker in *foma*[1], which includes an algorithm for spell checking called 'med search' (minimum edit distance search). Consequently, there is no need to include the edit operations in the implementation of the spell checker itself, 'med search' can be applied to a 'normal' finite state automaton. The *foma* spell checker performs its task much faster than the original *xfst* tool[2]: The correction of the word *wsaiyki* (correct *wasiyki*) with the *xfst* spell checker took 12 seconds on a dual-core AMD 64bit work station, whereas the foma version finds its suggestions in 0.2 seconds.

1.1. Characteristics of Southern Quechua

Quechua is a group of closely related languages, spoken by 8-10 million people in Peru, Bolivia, Ecuador, Southern Colombia and the North-West of Argentina. Ethnologue[3] also lists some Quechua speakers for Chile. The Quechuan languages are divided into two main branches, Quechua I and II in terms of the Peruvian linguist Torero. Quechua I is the more archaic group of dialects, spoken in Central Peru. It comprises a heavily fragmented dialect complex, with limited mutual comprehension between the different local varieties, although they share a number of clear common features (Adelaar and Muysken, 2004, 185). The origin of the Quechuan languages lies probably in this area (Cerrón-Palomino, 2003).

The second branch, Quechua II, comprises all the remaining Quechua dialects:

- QIIA, spoken in Northern Peru
- QIIB, spoken in Ecuador and Colombia
- QIIC, spoken in Southern Peru, Bolivia, and Argentina

The letters A-C stand for the linguistic distance to QI, QIIA is therefore the most akin to QI, whereas QIIC is the most divergent group respective to QI. As for our project, we focus on the Quechua IIC dialects, and within these especially on the Ayacucho and Cuzco variants. The reason for this choice is mainly due to non-linguistic circumstances: Quechua IIC is by far the best described dialect

[1] see `http://foma.sourceforge.net/dokuwiki/doku.php?id=start`

[2] The original *xfst* spell checker can be tested online at `http://kitt.cl.uzh.ch/kitt/quechua/quechua2.html`, the *foma* version can be downloaded at `http://kitt.cl.uzh.ch/kitt/quechua/download.html`

[3] `http://www.ethnologue.com`

group, and there are more bilingual texts available than for the other Quechua varieties.

We implemented two versions of the *xfst* spell checker, one for Cuzco and one for Ayacucho Quechua, whereas the current *foma* spell checker is meant to be used only with Cuzco Quechua. The division between Ayacucho and Cuzco Quechua is mainly due to the occurrence of glottalized and aspirated stops in the Cuzco (and Bolivian) dialects, a phonetical distinction absent in Ayacucho Quechua (Adelaar and Muysken, 2004, 187),(Cerrón-Palomino, 2003, 242-245).

2. Quechua Morphology

Quechua is a strongly agglutinative, suffixing language. There are more than 130 Quechua suffixes, the exact number, as well as the form of the suffixes exhibit substantial variation across dialects, even within the Quechua IIC subgroup. Available linguistic descriptions are not detailed enough for the implementation of a spell checker. Therefore, the morphotactical scheme on which our original analyzer and generator are based had to be carefully established by comparing different grammars, scanning large amounts of texts for suffix combinations and consulting with native speakers.

There are five functional classes of Quechua suffixes as described in Table 1. Besides the nominalizing and verbalizing suffixes, there are many nominal and verbal derivational, respectively inflectional suffixes. Additionally, Quechua has a small set of independent suffixes. These suffixes can be attached to both verbal or nominal forms, without altering the part of speech of the given word form. The position of these suffixes is at the end of the suffix sequence, their relative order is more or less fixed, though dialects show minor variations. The functions of the independent suffixes include data source, polar question marking and topic or contrast, amongst others. In combination with interrogative expressions, these suffixes may acquire special meanings (Adelaar and Muysken, 2004, 209). In combination with demonstrative pronouns, the independent suffixes may also take the place of conjunctions, which are virtually non-existent in Quechua, unless they are borrowed from Spanish (Adelaar and Muysken, 2004, 208).

The suffixes of the last three classes in Table 1 (nominal, verbal and independent suffixes) are grouped together in slots, a roughly simplified scheme of Quechua word formation is shown in Table 2.

There are two basic types of roots, the ones that take verbal suffixes, and the ones that take nominal suffixes. Every verbal root can be nominalized, but not every nominal root may form a derived verb, there are some restrictions, e.g. on personal pronouns. Generally, the nominalization and verbalization are extremely productive in Quechua word formation. Example 1 (*kachichasqa*) starts with a nominal root, gets verbalized and finally nominalized again. Example 2 (*yuyaychakusqaykikunawanmi*) on the other hand starts with a verbal root, gets nominalized, verbalized and finally nominalized again.

(1) *kachi* *-cha* *-sqa*
salt -Fact(VS) -Perf(NS)
'salted, salty'

(2) *yuya* *-y* *-cha* *-ku* *-sqa*
think -Inf(NS) -Fact(VS) -Rflx -Perf(NS)
-yki *-wan* *-mi*
-2.Sg.Poss -Inst -DirE
'with/by your thought'

(3) *chinka* *-y.*
loss/lose -1.Sg.Poss
'My loss.'

chinka *-ni.*
loss/lose -1.Sg.Subj
'I lose.'

As a matter of fact, a considerable number of Quechua roots are neither verbal nor nominal, but ambiguous: they can take nominal or verbal morphology without any derivation (see example 3). Additionally, there is a handful of particles that combine only with independent suffixes, e.g. *icha* - 'or'.

Quechua is for the most part an entirely regular agglutinative language. Nevertheless, there are some minor morphophonological features that have to be handled by special rules. There are roughly three cases of morphophonological changes when it comes to word formation: vowel deletion, vowel change and epenthesis.

Figure 1 gives a more detailed, yet still simplified overview of the spell checking finite state automaton: Some nominalizing suffixes yield converbs rather than nominal forms, those behave differently from other nominalized forms. Also, the independent suffix *-lla* has no fixed position, but may rather freely occur between the other suffixes. Furthermore, the independent suffixes at the end of the word form are split up into 7 different slots. These additional features have been omitted from Figure 1 due to lack of space.

As can be seen in Figure 1, a Quechua word form can start with different kinds of roots: There are personal (*Pers*), interrogative (*Intr*), and demonstrative pronouns (*Dem*) that may take the same suffixes as nominal roots (*NRoot*), with some restrictions. Additionally, there are two types of particles, some behave like nominals (*PrtV*), those particles may even be verbalized, while others can bear only independent suffixes (*Part*). There is only one kind of verbal root (*VRoot*). All nominal transitions may be empty (ϵ), as a consequence, a bare nominal root is a valid word form. In the verbal paradigm on the other hand, all transitions except slot 6, containing the person marker, may be empty: A verbal root alone makes no valid Quechua word form, at least a person marker is required. There are several transitions from the verbal to the nominal scheme via nominalizing and verbalizing suffixes. A further possibility for a bare nominal root is to be directly followed by a verbal root, this construction represents noun incorporation and is limited to unspecific objects of transitive verbs, e.g. *uywamichiy* - 'to herd animals' consisting of *uywa* - 'animal' and *michiy* - 'to herd'. Most people would write two words in this case, but some prefer the contracted version. Therefore, incorporation has to be considered for the implemen-

1	nominalizing	V → N
	llank'a -q	
	'work-Agentive' ⇒	worker
2	verbalizing	N → V
	wira -cha-	
	'fat-Factitive' ⇒	to grease
3	nominal	N → N
	wasi -su	
	'house-Augment.' ⇒	big house
4	verbal	V → V
	wañu -chi-	
	'die-Causative' ⇒	kill
5	independent	N → N
		V → V

Table 1: Suffix Classes

nominal root	verbal root
↓	↓
derivation	derivation
↓	↓
possession	object marker
↓	↓
number	aspect/tense
↓	↓
case	person
	↓
↓	modality
	↓
independent suffixes	

Table 2: Suffix Order

tation of a an analyzer.[4] Compounds of two nominal roots may also occur, e.g. *wawawasi* - 'day-care center, nursery' consisting of *wawa* -'child' and *wasi* - 'house'.

3. Orthography

Almost all native languages of the Americas struggle with strong social pressure from the dominant languages (mainly English, Spanish and Portuguese), a large number having become extinct already. The situation of Quechua is no exception: although its national varieties have been given the status of official languages in Peru, Bolivia and Ecuador, it is considered to be the language of the 'serranos', of the "country bumpkins", whereas Spanish is the language associated with education and modernity. Under these circumstances, parents are often more concerned about their children's Spanish skills than their competence in Quechua, as good knowledge of the Spanish language seems to be an essential prerequisite to climb the social ladder. Given this adverse situation, it is not surprising that only few people express themselves in written Quechua. An additional drawback is the lack of a widely accepted, standardized orthography: there are ongoing debates on this issue, and a common agreement does not seem to be within reach.

There are two major contrasts in Quechua IIC written texts. The first one is a purely dialectal divergence between the Cuzco/Bolivian dialects on one side, and the Ayacucho/Argentina varieties on the other side: Cuzco/Bolivian Quechua has, like Aymara, a three way distinction of stops (plain vs. glottalized vs. aspirated), whereas Ayacucho and Argentina Quechua have only simple stops. Whether an author writes the word for bread as *t'anta* or just *tanta* depends accordingly on the specific dialect he speaks.

The other point of controversy is entirely conventional. Quechua has three phonemic vowels: *a, i, u*. However, *e* and *o* occur as allophones in the proximity of post-velar *q*. While from a linguistic perspective it is evident that the writing of a word should consider only phonemes, not allophones, it seems that, probably due to the influence of Spanish orthography, a lot of people prefer to write *i* as *e*, respectively *u* as *o* according to pronunciation.

The writing of the Quechua vowels is subject of an ongoing debate concerning the elaboration of a written standard. While thoroughly rejected by linguists, the 5-vocalic spelling is strongly propagated by the Academia Mayor de la Lengua Quechua in Cuzco, alongside other peculiarities.

Obviously, the lack of a commonly accepted orthography is a major problem for spelling correction, as the decision whether a given word form is correctly written depends on the comparison with a 'gold standard'. As for now, we have implemented two spell checkers, one for the less controversial Ayacucho Quechua (using only plain stops), and one for Cuzco Quechua (including the distinction between plain, aspirated and glottalized stops). Both spell checkers use 3-vocalic spelling. For the Cuzco variant, the lexicon was built consulting the dictionary by the Academia, but using strictly the 3-vocalic writing and avoiding other oddities like the writing of aspirated *ph* as *f*.

As there are some confusions, the lexicon of our Cuzco dialect spell checker is currently being revised by a qualified native speaker.

4. Spell checking with *foma*

As mentioned in the introduction, we had already developed two finite state transducers, one for analysis and one for generation of Quechua word forms. Both tools rely on the same basic scheme of suffix combinations, but the analyzer is more tolerant with its input: different spellings, as well as divergent suffix forms and even dialectal variations in suffix order are recognized. As for generation, it is pointless to allow different orthographies or variation in suffix order, as this would only lead to multiple parallel output variants. The generator is therefore more restrictive and limited to Cuzco, respectively Ayacucho Quechua, whereas the analyzer is able to handle input from other Quechua IIC variants as well. Spell checking with *foma's* 'med search' requires a finite state automaton (Hulden,

[4] Actually, also the spell checker should be able to handle these forms, although it is not clear whether the contracted forms should be 'corrected' by a split or not.

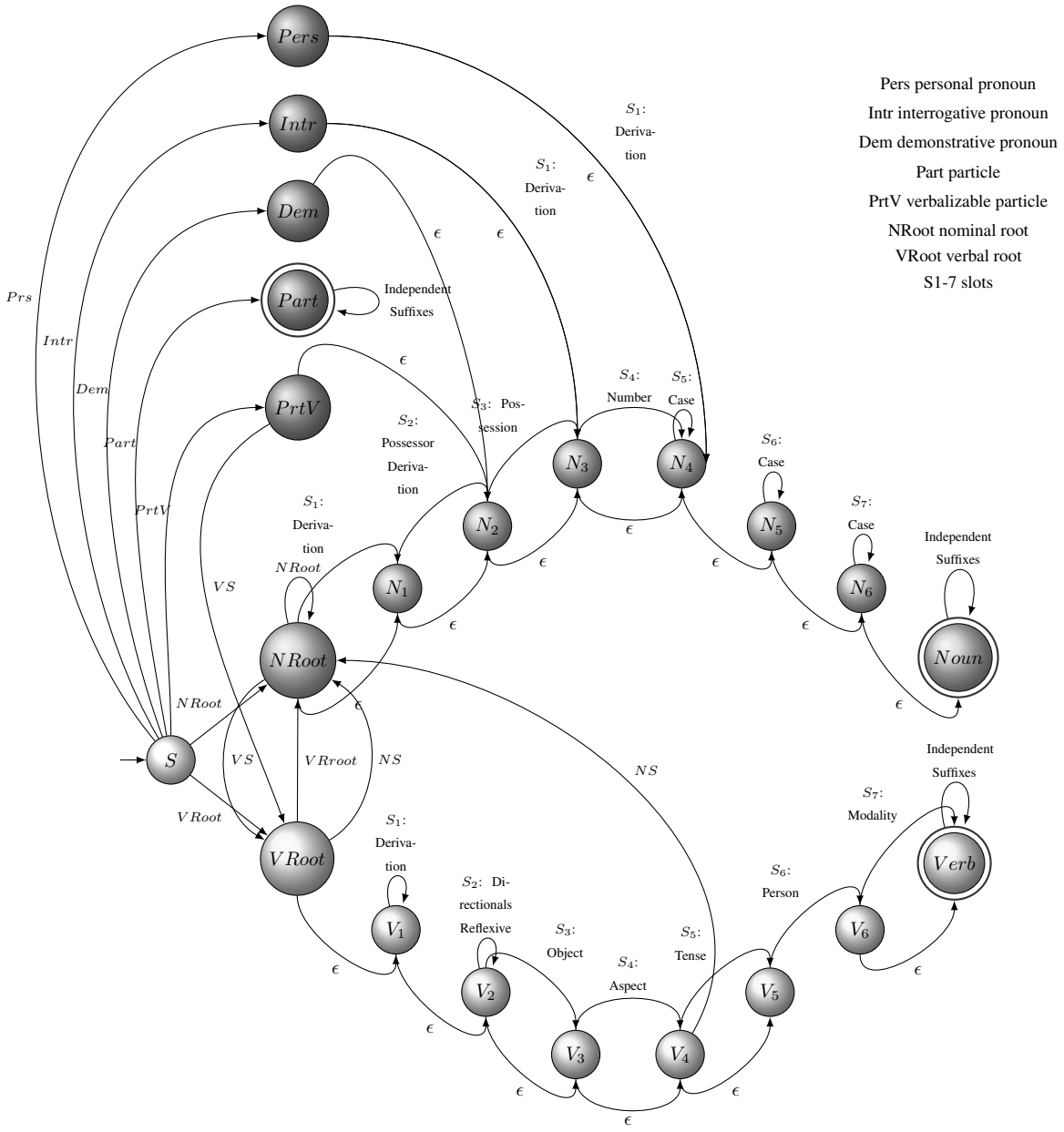

Figure 1: Simplified Quechua Finite State Automaton

2009). It is straightforward to use the lower side (natural language side) of the generation transducer for this purpose.

foma's 'med search' calculates the minimum deviation of a given input string from the recognized, i.e. correct, strings of the regular language implemented by the automaton. The error metric in 'med search' is the same as the one used in the original *xfst* spell checker: the Levenshtein distance, calculated by counting the basic edit operations (insertion, deletion, substitution) necessary to convert one string into another. The possibility to adjust the Levenshtein distance ('cost') for specific, language-dependent edit operations through so-called confusion matrices is a further benefit. As for Quechua, the cost of substitutions of *i* with *e* and *u* with *o* should be lower than other substitutions, see the example confusion matrix for Quechua in Figure 2: This confusion matrix states that substitutions of *i* with *e*, resp. *u* with *o*, can occur at cost 0, while the cost for substitutions of other letters is 2. Figure 3 shows the output of 'med search' applied to the word *orqo*, correctly spelled *urqu* - 'mountain'. Considering only edit distances, *orqo* is two substitutions away from *urqu*, yet for Quechua, the confusion matrix states that substitution of *u* with *o* may occur at zero cost. This assures that in cases of 5-vocalic writing the corresponding 3-vocalic spelling will occur on top of list of suggestions, as otherwise correctly written words will have zero cost, as it is the case with *urqu*. The following suggestions (*urquy*, *urqus* and *urqun*) are considered worse, as they all require the insertion of an additional letter, their cost is 1. The second example shows spell checking of the word *kachichasqa* (see example 1 in section 2), misspelled as *kachichasa*. There are more than one word forms at Levenshtein distance 1, which all appear in the

```
Insert 1
Substitute 2
Delete 1
Cost 0
i:e u:o
```

Figure 2: Confusion Matrix Example

Example 1		Example 2	
input:	orqo	input:	kachichasa
correct:	urqu	correct:	kachichasqa
suggestions:	cost:	suggestions:	cost:
urqu	0	kachichas	1
urquy	1	kachichasqa	1
urqus	1	kachichaspa	1
urqun	1	kachichasaq	1
		kachichasá	2

Figure 3: Spell Check Examples

list of suggestions. The maximum edit distance, as well as the maximum number of suggestions may be changed in *foma* through the setting of the global variables 'med-cutoff' and 'med-limit'.

The size of the *foma* spell checker is 5.1 MB and its lexicon contains more than 3000 roots. A major disadvantage for spell checking is that 'med search' is only applicable from the *foma* interface. Therefore, we implemented an additional utility called 'fmed', that enables spell checking directly from the shell (analogous to the 'flookup' utility already contained in *foma*).

5. Conclusions

We implemented a basic spell checker for Cuzco Quechua based on the error metric of Levenshtein distance. A more sophisticated approach would take morphotactical errors into account, e.g. the suffix of direct evidence has the form *-mi* after consonants and *-n* after vowels. A word form like $^{*}wasimi$ (*wasi* - 'house') should be corrected directly to *wasin*. Yet, the Levenshtein distance between these word forms is 3: Deletion of one character at cost 1, and substitution of another character at cost 2. On the other hand, the completely different word *simi* - 'mouth' is closer, at only Levenshtein distance 2 (deletion of two characters). As a consequence, the useless suggestion *simi* will be ranked above the intended correct word form *wasin*.

A similar problem is the common influence of Spanish orthography in the writing of Quechua words, e.g. *k* is often written as *qu*, like in $^{*}purinqui$ instead of *purinki* - 'you walk'. As above, unwanted suggestions as *purini* - 'I walk' have a better Levenshtein distance score than the intended correct word form. It would be practical to have a confusion matrix not only at the level of symbols, but of strings.

Another solution is to connect several finite state automata instead of using only one. We have used this strategy for the original *xfst* spell checker. It consists of several cascaded automata: the first one recognizes only correct word forms, the second recognizes morphotactical errors, and the last two handle word forms at Levenshtein distance 1 and 2. Due to this setup, the original *xfst* spell checker handles morphotactical errors correctly. A similar setup could be done with the foma spell checker.

An important issue that needs to be addressed is the handling of Spanish loan words. Almost every Quechua text contains a large number of Spanish loan and foreign words. The difference between the two categories is that while loan words are used as verbal or nominal roots with suffixes directly attached to them, foreign words take no suffixes but instead are cited with *nisqa* - 'said, called': *nisqa* then bears all the corresponding suffixes. Loan words are often adapted to Quechua pronunciation in their spelling, but there is absolutely no standard on the spelling of Spanish loans, e.g. Spanish *república* - 'republic' might be written as *republica, ripuwlika, ripublika* or even with the original Spanish spelling. Foreign words on the other hand usually maintain their original spelling. While the analyzer is able to handle a limited number of loan words contained in an additional lexicon file, the spell checker has no module so far to deal with words of Spanish origin. We also plan to implement a suitable interface, e.g. as web application, as the output of *foma's* 'med search' is not user friendly. An integration into open source text processing systems like Libre Office would be a further enhancement.

6. References

Willem F. H. Adelaar and Pieter Muysken. 2004. *The Languages of the Andes*. Cambridge Language Surveys. Cambridge University Press.

Iñaki Alegria, Maxux Aranzabe, Nerea Ezeiza, Aitzol Ezeiza, and Ruben Urizar. 2002. Using Finite State Technology in Natural Language Processing of Basque. In *CIAA '01 Revised Papers from the 6th International Conference on Implementation and Application of Automata*. Springer.

Rodolfo Cerrón-Palomino. 2003. *Lingüística Quechua*. Centro de Estudios Regionales Andinos Bartolomé de Las Casas (CBC), 2. edition.

Mans Hulden. 2009. Fast approximate string matching with finite automata. *Procesamiento del Lenguaje Natural*, (43):57–64.

Kemal Oflazer. 1996. Error-tolerant Finite State Recognition with Applications to Morphological Analysis and Spelling Correction. *Computational Linguistics*, 22(1).

Tommi Pirinen and Krister Lindén. 2010. Finite-State Spell-Checking with Weighted Language and Error Models. Building and Evaluating Spell-Checkers with Wikipedia as Corpus. In *Proceedings of LREC 2010 Workshop on Creation and use of basic lexical resources for less-resourced languages*, May.